DOCUMENT RESUME

ED 453 275 TM 032 812

AUTHOR Skaggs, Gary; Tessema, Aster
TITLE Item Disordinality with the Bookmark Standard Setting Procedure.
PUB DATE 2001-04-11
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adults; Difficulty Level; English (Second Language); *Language Tests; Sampling; Standards; Test Construction; *Test Items
IDENTIFIERS Rasch Model; *Standard Setting

ABSTRACT

This paper presents an application of the bookmark procedure to a test comprised of increasing text difficulty levels. The Test of English Proficiency for Adults (TEPA) was used for this study. Three forms of the TEPA were field tested in 1999 with approximately 1,000 non-native English speaking students enrolled in English-as-a-Second-Language programs. Bookmarking was generally successful with this test, but item disordinality played a major role in discussions by the judges. Three possible explanations were considered. The findings suggest that some degree of item disordinality resulted from sampling error of the item parameters. Basing the item map on the three-parameter model as opposed to the Rasch model had little impact on item placements in the booklets. An exploration of text difficulty suggested that the judges did not base their bookmark placements solely on text difficulty. The greatest consistency was found among the placements for items associated with the same text difficulty level. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

G. Skaggs

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 453 275

Item Disordinality with the Bookmark Standard Setting Procedure

Gary Skaggs
Virginia Polytechnic Institute & State University

Aster Tessema
GED Testing Service

Abstract

In many applications of the bookmark standard setting procedure, discussions of item disordinality arise that can ultimately affect where the judges place their bookmarks. This paper presents an application of the bookmark procedure to a test comprised of increasing text difficulty levels. Bookmarking was generally successful with this test, but item disordinality played a major role in discussions by the judges. We examined three possible explanations. Our findings suggest that some degree of item disordinality resulted from sampling error of the item parameters. Basing the item map on the three-parameter model as opposed to the Rasch model had little impact on item placements in the booklets. An exploration of text difficulty levels suggested that the judges did not base their bookmark placements solely on text difficulty. The greatest consistency was found among the placements for items associated with the same text difficulty level.

TM032812

Paper presented at the 2001 Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

BEST COPY AVAILABLE

Item Disordinality with the Bookmark Standard Setting Procedure

Introduction

Standard setting has become an integral component of the current standards-based reform movement in American education. Many methods have been proposed and used to set standards on tests, and there is still a great deal of controversy over their use. One method that has risen fast in popularity is the Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996; Lewis, Green, Mitzel, Baum, & Patz, 1999). It has been used successfully in a number of state and local school district assessments and in different subject areas

The bookmark method was originally developed to respond to problems that have been observed with the widely used modified Angoff standard setting procedure (Angoff, 1971). First, the Angoff method requires panelists to make a judgment about each item in the test, a task that is both labor intensive and cognitively complex. The bookmark method requires only a single judgment about a collection of items. Second, bookmarking accommodates multiple item formats in a single standard setting, while the Angoff method was intended for multiple-choice items. Modifications of the Angoff method for polytomously scored items tend to result in different standards for each item format (Pellegrino, Jones, & Mitchell, 1999).

The essential idea of the bookmarking method is to use an item response theory (IRT) model to develop an item map with the items arranged in difficulty according to a response probability (rp) criterion. Huynh (1998) recommended using an rp criterion that maximizes the item information function for a correct response, which is .67 for the Rasch and two-parameter models and (2+c)/3 for the three-parameter model (where c is the lower asymptote parameter). This rp criterion is widely used in bookmark implementations and promoted by Lewis et al. (1999).

In a typical bookmark standard setting meeting, judges develop a description of what knowledge, skills, and abilities should be evident at the standard in question. These performance level descriptions are then applied to a booklet of test items that is constructed from an item map. The items in the booklet are arranged in order of increasing difficulty at the .67 probability of a correct response. Typically, there is one item per page. The task of each judge is to determine the item in the booklet such that examinees who meet the standard should be able to answer items correctly up to and including that item but not be able to answer subsequent items, with a .67 probability. For further details concerning the implementation of the bookmark, the reader is referred to Lewis et al. (references listed above).

The purpose of this paper is two-fold: 1) to report the use of the bookmark method on a type of test for which the method has not yet been applied (to the best of our knowledge), and 2) to explore a problem in the above application, namely item disordinality, or the disagreement among the judges on the ordering of the items in the booklet.

Using the Bookmark Method for a Test of English Language Proficiency

Test Design

The Test of English Proficiency for Adults (TEPA) was developed by Second Language Testing, Inc., and the GED Testing Service to assesses the English reading skills of non-native adults. The TEPA was designed to serve as an adjunct to the Tests of General Educational Development (GED) battery when an examinee takes the GED Test in a language other than English. The TEPA could also stand alone as an instrument for placement or promotion in adult ESL programs and community colleges, or for use by employers who need to know the reading proficiency of non-native adults before hiring or promoting them.

The TEPA consists of text passages of varying difficulty levels accompanied by a short series of multiple choice items. The texts themselves are defined at one of six levels of reading difficulty, from an adult basic education "high beginner" through the freshman year of college. The intent of the score on the TEPA is to locate the highest level of text difficulty at which the examinee can read English proficiently. The texts are authentic examples of general English, drawn from advertisements, forms, newspapers, and other daily reading material. On the TEPA forms, texts are presented in order of difficulty, so that an examinee responds to items based on two to three texts at the first level, then to items based on two to three texts at the second level, and so forth.

Three forms of the TEPA were field tested in the summer of 1999 with approximately 1,000 non-native English speaking students enrolled in ESL programs at community colleges, adult education programs, and inmates taking ESL classes in federal prisons. Each of the three TEPA forms field tested had 72-78 items. This sample was diverse and included individuals at all proficiency levels.

Bookmarking

Classical and IRT item and test analyses were carried out on the field test data. Due to the relatively small sample size (about 350 per form), the Rasch model was used to calibrate item difficulties from the three forms. The BILOG 3 (Mislevy & Bock, 1989) program was used for item calibration. Common items among the three forms were used to place the item difficulties from the three forms on a common scale.

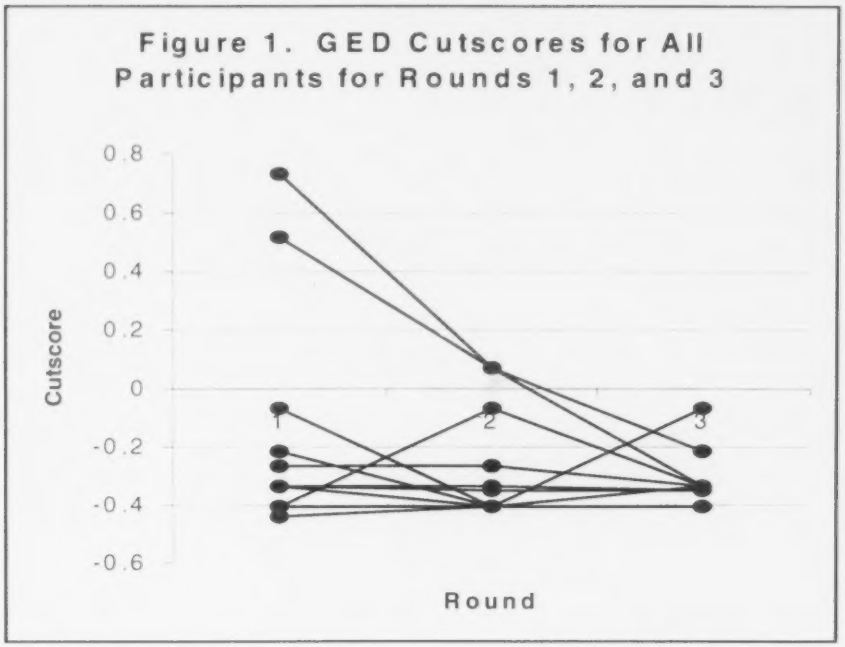
As a result of these analyses, 103 items from the three field test forms were selected for potential inclusion in the final operational form. This set of 103 items was assembled into a binder in ascending order of difficulty, with text on the left-hand page and items on the right-hand page. The texts and items appeared as they had in the field test forms, except that the items were re-numbered to reflect their new order in the book, and only one item appeared with each text. When several items from a particular text were included in the binder, the text appeared again with each item.

Fifteen judges were selected for the panel from a variety of backgrounds and occupations, all with some degree of experience in teaching, testing or credentialing the TEPA's target population. Six were teachers or administrators from adult ESL programs. Six others were employed by a state department of education. There was one independent consultant (a specialist in vocational ESL), one employee of a local school district, and one educational researcher from another testing organization. All judges had relevant skills and experience, permitting them to make informed judgements regarding necessary skills for examinees and the needs and concerns of society, industry, and state education agencies.

The standard setting took place in December 2000. Procedures recommended by Lewis et al. (1999) were followed. The first part of the meeting centered on developing a performance level description for a passing score. Each judge commented on a population of non-native adults with which he or she was familiar. Judges mentioned the following text types as among those that the minimally competent examinee should be able to read and understand: common forms, bills, safety warnings and procedures, medicine labels, transportation schedules, children's report cards, tables of contents, alphabetized indexes, traffic signs, form letters, and some simpler stories, newspaper articles, and tax forms. At the conclusion of this process, a rough consensus emerged about what reading skills the minimally competent examinee should have in order to be awarded a GED credential.

To set the standard for the GED credential, the panel of judges and staff members broke into three groups of five judges each, with a psychometrician and a content specialist from GED staff present at each table to moderate and address any questions that arose during the discussion. Three rounds of bookmarking took place. In the first round, each judge read individually through a binder of items, placing bookmarks. Then, the judges explained why they placed bookmarks where they did. Other members of the group were invited to disagree, if they felt the placement was inappropriate. Following this discussion, judges placed their bookmarks a second time. For the third round, all three groups convened as one large group. Each judge shared his or her bookmark placement from the second round with the large group, followed by further discussion of the rationale for particular bookmark placements. The judges then placed their bookmarks a third and final time. A final cut score was calculated as the median of the individual judges' cut scores from the third round.

Figure 1 below shows the locations of judges on the three rounds in terms of the logit locations of the items. Each line represents a judge. Since several judges had identical ratings in all three rounds, their lines lie atop one another and appear as one line. Clearly, greater consensus was achieved at each round. In addition, the consensus was a downward trend. That is, the judges with the highest cut scores at round 1 tended to lower their cut scores, while the judges with the lowest cut scores usually remained the same.



Although the use of the bookmark method was considered a success, one issue that affected the judges' placement of the bookmarks was disagreement on the ordering of the items within the booklet. Several judges insisted that several items should have been placed before other items at different locations in the booklet. In some cases, the judges' locations for the items fell at significantly different places in the booklet.

The effect of this disagreement can be seen in Figure 1. The two highest judges placed their bookmarks where they did largely because they perceived that items near the end of the booklet were easier than their empirical difficulties indicated. Other judges changed their bookmarks back and forth between two items that they perceived to be about equal in difficulty, but which were separated by several intervening items.

Exploring the Effect of IRT Model on Item Disordinality

Lewis and Green (1997) addressed item disordinality as an issue that has appeared in virtually all applications of the bookmark method. They offer two explanations: 1) local curricular differences among the judges, and 2) the inability of judges to estimate item difficulty accurately. As a solution, Lewis and Green recommend a detailed discussion among judges of each item as to what it measures and why it is more difficult than the preceding item. Using this approach, Lewis and Green find that most disordinality disagreements are resolved.

In this application, however, such discussions did not completely resolve the disagreement. Three explanations for this seem possible. First, there is obviously sampling error in the estimation of all item parameters, and the item ordering could change to some extent just from this source of error. A standard error of Rasch item difficulty of .15 (a typical value in this study) would correspond to an approximate 95 percent confidence interval of plus or minus .30. Most of the logit locations (i.e. theta such that P=.67) differed by less than .05, in some cases by less than .01. At the point in the map where the standard was ultimately set, a difference of plus or minus .3 logits could have resulted in a change of about 10 locations in either direction. Therefore, random sampling error could have accounted for some of the perceived item disordinality.

In Figure 1, sampling error could account for several judges' going back and forth between two items that were fairly close together in the booklet. When discussing item location differences of 10 places or less, the judges should probably be advised that these differences are not significantly different.

Second, the judges may have ignored characteristics of items that affect difficulty, such as guessing, the quality of distractors, and item discrimination. Some support for this hypothesis has been reported by Shepherd (1995), who summarized the findings of National Academy of Education's evaluation of standard setting procedures used for the National Assessment of Educational Progress. This evaluation found that judges were unable to take guessing into account when judging the difficulty of items (p. 151)

A third possible reason for disagreement in the perceived item placement is that the performance level descriptions for passing the TEPA combine the level of difficulty of the passage with skills related to what examinees should be able with those passages. Most of the disagreements over item order stemmed from items pertaining to relatively easy texts being placed after items from more difficult texts.

The next two sections of the paper explore these last two possibilities in terms of the effect of IRT model on item disordinality and a consideration of text levels in bookmarking.

Method

Parameters for all items on the three field test forms were re-estimated using the three-parameter model with BILOG 3. Admittedly, the sample sizes (about 350 per form) were considerably lower than are usually recommended. However, because both item difficulties and examinee abilities spanned a wide range, we felt that BILOG would be able to provide reasonable enough estimates for this exploratory study.

By using prior distributions for all three item parameters, BILOG easily converged. Standard errors for the "a" parameter averaged about .30. Standard errors for the" b "parameter averaged about .30 as well, compared to about .15 using the Rasch model. Clearly, item difficulties using the Rasch model had less sampling variance than their counterparts using the 3PL model. On the other hand, the items used in the item map were the highest quality items from the field test. Thus, items were excluded that were extremely easy or difficult, had weak discriminations, or had weak or problematic distractors. These also tended to be the items with the highest standard errors for their estimated item parameters.

Using the BILOG item parameter estimates, the item map was then reconstructed. Item parameters from the three field test forms were placed on the same scale. This was accomplished using linear transformations described in Chapter 1 of the BILOG manual and based on the overlapping items. That is, item parameters were transformed so that the item difficulties for the common items had equal means and that their item discriminations had equal geometric means. Once this was accomplished, the logit locations for each item were determined using the correction for guessing as described by Lewis et al. (1999).

Results

Figures 2 and 3 below show the item characteristic curves (ICCs) for the item maps based on the Rasch and 3PL models, respectively. Two things stand out from these figures. First, in both figures, the ICCs are very close together relative to the logit scale. In other words, a small increment in theta in either direction results in crossing a large number of ICCs. Second, due to unequal item discriminations, a number of ICCs cross with the 3PL model.

Figure 2. ICCs for Items in Item Map Based on the Rasch Model

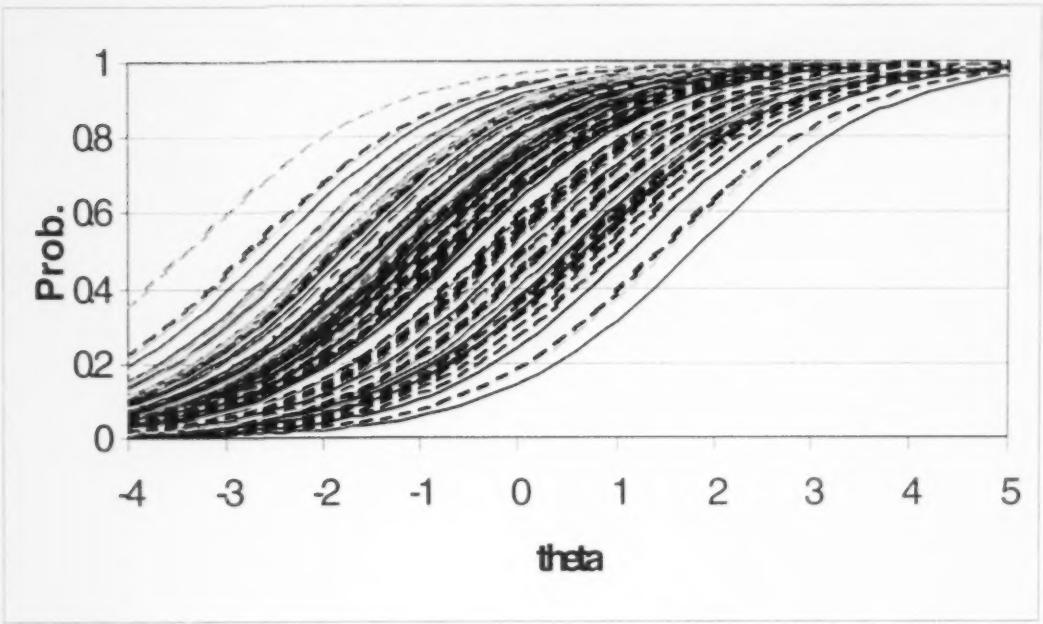
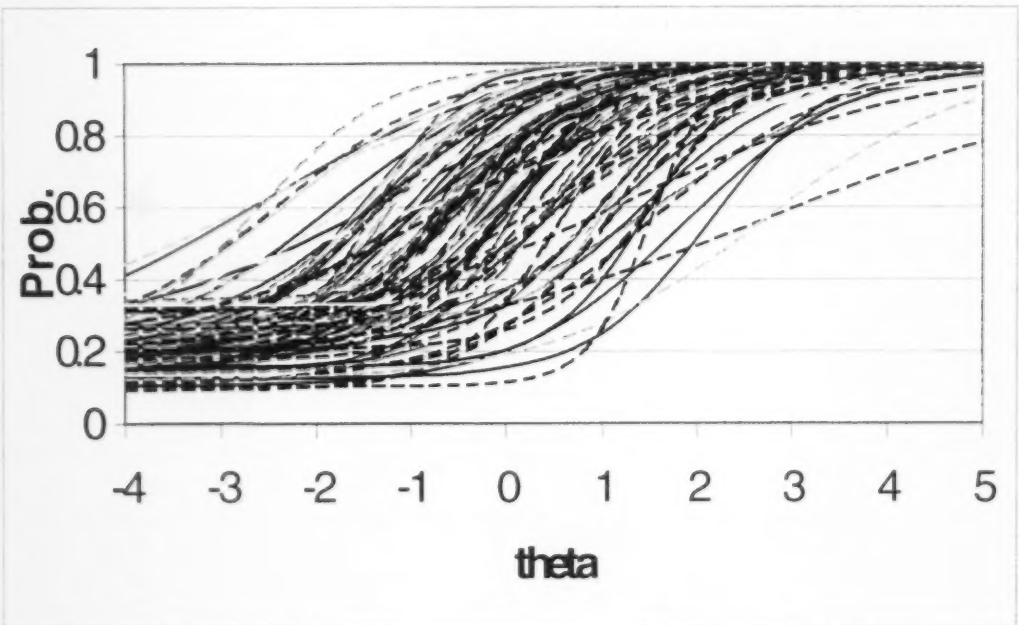


Figure 3. ICCs for Items in Item Map Based on the 3PL Model



However, the logit values corresponding to probabilities of .67 with the Rasch model and .67 (corrected for guessing) with the 3PL model correlated at .967. The rank-order correlation between the two sets of item placements was .985. In terms of actual differences in item placement (i.e. from 1 to 103) with the booklet, the mean of the absolute value of differences in placements was 3.9, meaning that, on average, items were about 4 places apart between the two item maps. The largest difference was 16. For 48 items--nearly half--placement on the 3PL item map differed by 2 places or fewer from their placement on the Rasch item map. These differences lie within the approximate confidence interval discussed above. Yet the larger disagreements for the judges lay outside this interval.

Another way to view the comparison between models is to estimate how the three rounds of bookmarking would have gone if the 3PL item map and booklet had been used. Admittedly, it is impossible to know for sure how the judges' bookmarks might have been placed differently, given an alternate ordering of items. Presumably, a discussion of item disordinality would have taken place that might have altered their placements. However, if the judges had placed their bookmarks in the 3PL booklets at the same item locations as they did on the Rasch-based booklet, would the variation between judges have decreased?

Figures 4 and 5 below show the three rounds of bookmarks according to item locations within the booklet (i.e. from 1 to 103) for the Rasch and 3PL booklets, respectively. Figure 4 mirrors Figure 1, except that item location rather than logit value appears on the vertical axis. The figures are very similar, reinforcing the earlier conclusion that the choice of model made little difference in the convergence of bookmark placements.

Figure 4. Judges' Bookmark Placements for the Rasch-based Booklet

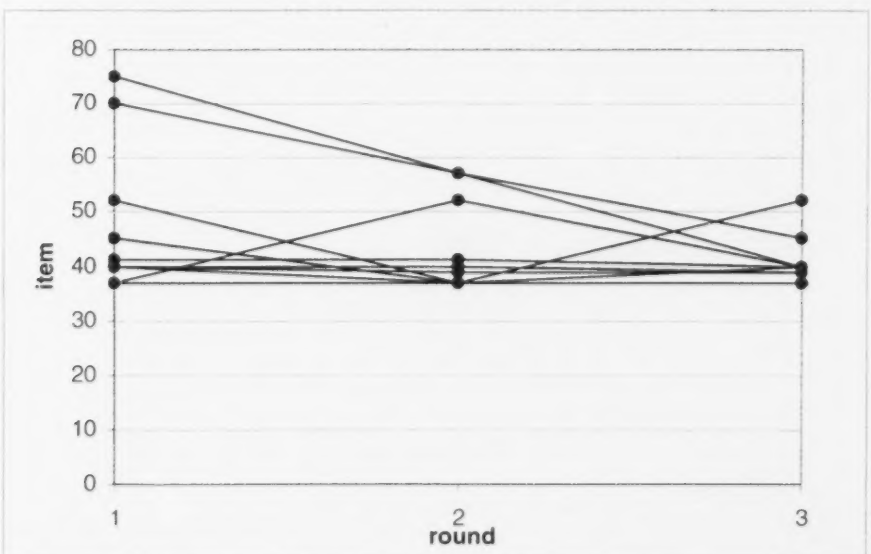
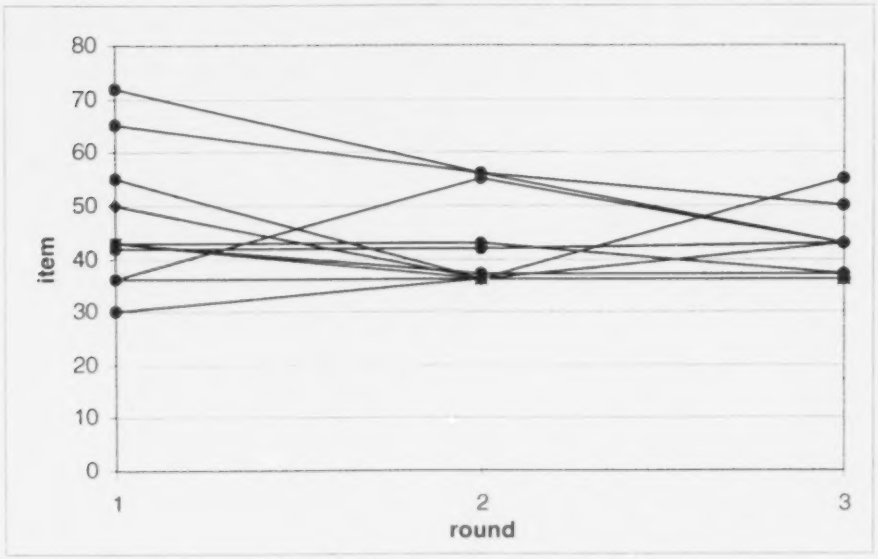


Figure 5. Judges' Bookmark Placements for the 3PL-based Booklet



Finally, Table 1 below shows descriptive statistics for the judges' bookmark placements. For both models, the median item locations in the second round were the lowest of the three. However, the final median location for the Rasch model was item 39, only one item removed from the median in the first round. For the 3PL model, the final median was five items lower than that for the first round. In terms of variability of the judges' ratings, the standard deviations and ranges for the Rasch model were less than those for the 3PL model.

Table 1. Descriptive Statistics for Judges' Bookmark Placements (Item Locations)

		Round		
	Model	**1**	**2**	**3**
Median	Rasch	40	37	39
	3PL	42	36	37
St. Dev.	Rasch	12.3	7.5	4.1
	3PL	11.9	8.0	5.9
Range	Rasch	31	21	16
	3PL	43	21	20

The results of this analysis suggest that changing from the Rasch model to the 3PL model would not have significantly changed the placement of items in the item map and,

consequently, placement of bookmarks in the booklets by the judges. Therefore, a change to the 3PL model would not by itself have reduced the item disordinality among judges. Furthermore, if the judges had placed their bookmarks at the same item in each booklet, then the 3PL model would have produced slightly more variability between judges.

Verification with an Operational Test Form

As noted above, we had some concern about applying the 3PL model to such a small data set. We therefore carried out the above analysis with a large, nationally representative sample responding to an intact, operational test form. We used a form of the GED Test 4: Interpreting Literature and the Arts. This test consists of fiction, poetry, and drama texts with a set of multiple-choice items associated with each text. The data came from the norming sample for this test (a nationally representative sample of about 900 graduating high school seniors).

Item maps for this were created in the same manner as the TEPA using BILOG3 to estimate item parameters for the Rasch and 3PL models. The item positions on the two item maps were then compared. The item locations rank order correlation was .971, and the correlation between logit values (at the rp criterion) was .989. The mean of the absolute value of the difference between item positions was 2.25, with the largest difference being 6 positions. With this data set, therefore, there was less difference between the two item maps than with the TEPA data set, but in both cases, there was very little practical difference between using item maps based on the Rasch and 3PL models.

Consideration of Text Levels

The TEPA texts were classified by the test developers according to six levels. Table 2 below provides brief descriptions of the levels. Three basic factors underlie the difficulty levels of the texts: 1) complexity of sentence structure, 2) level of abstraction of the concepts in the texts, and 3) level of vocabulary. Because all three factors differ at each level, texts at each level were easily distinguishable by the judges. When the performance level description for passing is defined in terms of what types of texts examinees should be able to read, judges can readily identify the types of texts to which the description applies.

Table 2. Descriptions of Text Levels in the TEPA

Level	Description
1	Short, simple sentences giving information in the context of daily life; vocabulary limited to basic words used commonly in speech; accompanying pictures are often necessary.
2	Mostly short, simple sentences, but some more complex; texts refer to the broader world of adult life; vocabulary is mostly basic and concrete; some pictures.
3	Some simple, some complex sentences; texts refer to daily life, plus concepts learned more formally at work or school; vocabulary broader than ordinary speech.
4	Sentences can be long, complex, and detailed; texts refer to social and work life, but also abstract concepts or technical terms not defined in the passage; vocabulary mostly familiar but outside the bounds of ordinary speech.
5	Broad range of complex sentence structures; texts contain frequent references to abstract ideas not explained within the passage; vocabulary can include less common words, idioms, and professional jargon.
6	Sentences can be long, complex, and formal in style; texts contain a wide variety of unfamiliar subjects and can assume some ability to engage in theoretical thought; vocabulary can include many long, uncommon, and technical words.

By contrast, the performance level description of those skills examinees should be able to demonstrate with the target level of text may have been less clear. The items that were associated with each text level spanned a range of difficulty. These ranges overlapped across text levels. Figure 6 below shows a box-plot of the logit values (at P=.67) for each level of text. The overall level of item difficulty increased with text level. However, there was considerable overlap. For example, some items in the item map accompanying level-2 texts were more difficult than items accompanying level-4 texts. It is quite possible that despite our best efforts, the judges were unable to assess consistently either the skills measured by items or the level of reading required by the texts.

Figure 7 below shows judges' bookmark placements, according to the level of the text corresponding to the items at which they placed their bookmarks for the three rounds. This figure shows less convergence than the above figures for item placement, and suggests that judges did not base their bookmark placements entirely on the level of text difficulty.

Figure 6. Boxplot Showing the Range of Item Logit Values by Text Level

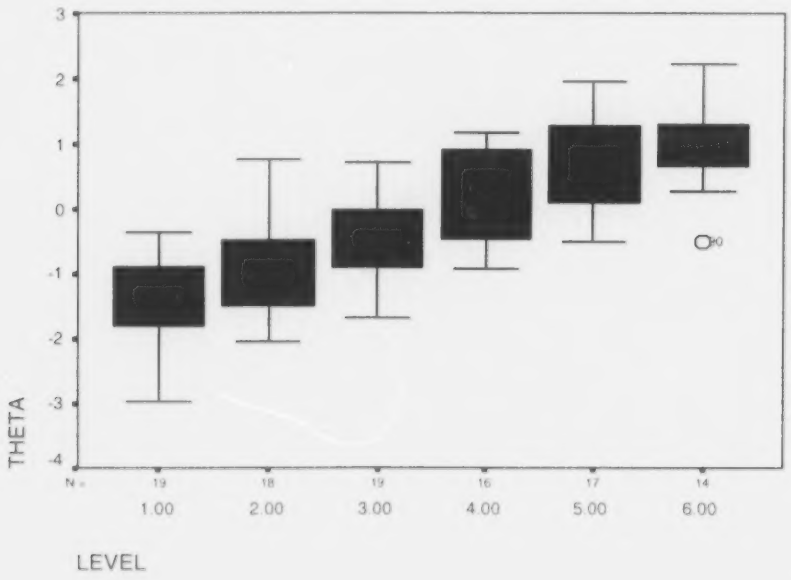
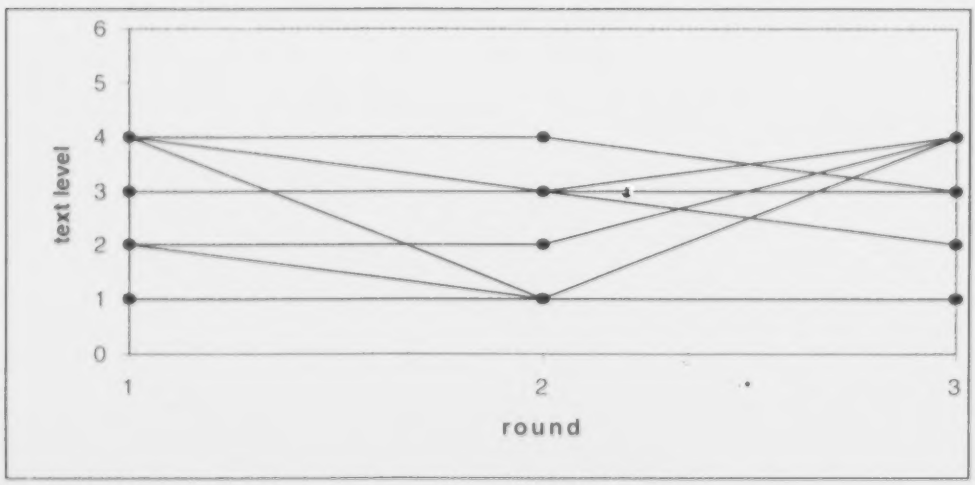


Figure 7. Judges' Bookmark Placements according to Text Level



Another way of viewing the effect of text level in the bookmark is to examine the judges' placements separately within each text level. This was done for each text level. For example, for all items associated with level 1 texts, where did the judges place their bookmarks?. The results are shown in Table 3 through 6 below. There are no tables for

level-5 and level-6 texts because no bookmarks were placed at items associated with texts at these levels. For level-1 texts, all of the placements were located at the highest end of the logit scale, indicating a ceiling effect for items associated with this level of text. On the other hand, the placements for level-4 text items showed somewhat of a basal effect. Relatively few of the bookmarks were placed at items with level-2 or level-3 texts. Those for level 2 were located in the middle of the logit range. The placements for level-3 text items showed the most amount of variability. They tended to be placed at lower logit levels in each succeeding round, but since there were few placements, this is a very cautious finding. Whatever the text level, the logit values where bookmark placements were made tended to be similar.

Table 3. Bookmark Placements for Items Associated with Level 1 Texts

	Round		
Logit	**1**	**2**	**3**
-0.360	*	*	*
-0.697	******	*********	*******
-0.731	*		
-0.822			
-0.832			
-0.964			
-1.063			
-1.174			
-1.286			
-1.467			
-1.541			
-1.694			
-1.721			
-1.759			
-1.844			
-2.130			
-2.275			
-2.338			
-2.972			

Table 4. Bookmark Placements for Items Associated with Level 2 Texts

	Round		
Logit	**1**	**2**	**3**
0.765			
-0.018			
-0.190			
-0.341			
-0.483			
-0.508			*
-0.537	*		
-0.558	*	*	
-0.706			
-1.164			
-1.374			
-1.397			
-1.419			
-1.492			
-1.549			
-1.758			
-1.873			
-2.055			

Table 5. Bookmark Placements for Items Associated with Level 3 Texts

	Round		
Logit	**1**	**2**	**3**
0.709			
0.225	*		
0.114			
0.110			
0.085			
-0.110			
-0.222		**	
-0.346			
-0.355			
-0.368			
-0.641		*	**
-0.732			
-0.755			
-0.797			
-1.004			
-1.066			
-1.209			
-1.430			
-1.684			

Table 6. Bookmark Placements for Items Associated with Level 4 Texts

	Round		
Logit	**1**	**2**	**3**
1.164			
1.128			
1.030			
0.944			
0.843			
0.442	*		
0.323			
0.170			
-0.114			
-0.383			
-0.393			
-0.401			
-0.509			
-0.628	*	*	****
-0.683			
-0.925			

Discussion

Because of its limitations, this study must be considered highly exploratory and speculative. The sample size for item parameter estimation was small for using IRT. In addition, the sample itself was a convenience sample from the potential test user population. In addition to sampling issues, the above analysis of the 3PL model centered on "what if " cases: that is, had the judges picked the same items in a booklet of items in a different order than for the Rasch model, and thus ignored any context effects. In an application of the bookmark method, these context effects could be significant owing to the discussion about why the items were ordered the way they were. Finally, the item maps were based on selecting items from three different field test forms. Data from an operational form of these items might produce an item map and ordering quite different from that found here. We make no attempt here either to minimize these limitations or to argue that they had no relevance. Instead, our findings suggest further areas of inquiry under more carefully controlled conditions.

This application of the bookmark method was generally successful for a test of English language proficiency, organized around texts of increasing levels of difficulty. Greater consensus was achieved among the judges with each round of bookmarking. The item difficulties generally increased with text level, but there was considerable overlap. This meant that the items in the item booklet were not directly ordered according to text difficulty. In turn, this resulted in a lengthy discussion of what judges perceived to be a more correct ordering of the items. When placing their bookmarks, several judges

appeared confused over this ordering, and it affected their placements and increased their variability. The issue of item disordinality has been an issue in other bookmark applications, but we suspect that the problem may have been more severe in this type of test.

This study explored three possible explanations for item disordinality: 1) sampling fluctuation in the estimation of item parameters, 2) specification of a more highly parameterized model, and 3) the combination of text difficulty and item skills in the performance level description for the cut score.

For the TEPA item map using the Rasch model, an approximate 95 percent confidence interval around an item's difficulty was plus or minus .30. Most of the differences between adjacent items were in the .01 to .05 range. This meant that item location differences of up to 10 places were not meaningfully different. Some of the disordinality discussions at this standard setting were in this range. Rather than have judges try to agree on these item locations, it would probably have been wiser to point out that these items could easily change places with a different sample.

Our follow-up analysis with an operational test form and with a larger and more carefully designed sample resulted in a smaller confidence interval for sampling error. Our best advice would be to determine this confidence interval for each application and convey this interval to the judges.

Our exploration of model specification suggests that it is unrealistic to expect a more fully parameterized model to result in an alternative ordering items that will reduce item disordinality, as perceived by the judges. With the 3PL model, the logit values of the items (at the rp criterion of .67, corrected for guessing) were rank-ordered in nearly the same way as with the Rasch model. This finding was corroborated with a large, nationally representative sample on an operational test form.

Another possible explanation for the judges' perceived item disordinality was that the performance level descriptions for this test were based on the level of text difficulty, as well as on the skills pertaining to texts. Most of the discussion about disordinality centered around which case represented a higher level of performance: answering relatively easy items from more difficult texts or answering relatively difficult items from less difficult texts. Several judges appeared to bounce back and forth on this dilemma in their bookmark placements. Our study suggests that the judges in this bookmark application relied solely on neither text difficulty nor item difficulty.

Finally, in order to circumvent the problem of text level versus item level, we suggest one possible solution. First, items from the same level of text difficulty could be grouped together. Judges would then place bookmarks separately for each text level. A final cut score could be a function of the logit cut scores at each text level (such as a median, or some point between ceiling and basal levels). Essentially, this procedure would control for text level.

Had we been able to present separate booklets for each text level at the standard setting meeting, it is quite possible that item disordinality would not have emerged as an issue in the discussion of the performance level descriptions. In this study, when items at the first level of text difficulty were isolated, a ceiling effect for the item logit values was observed. Likewise, a basal effect was observed for items at the fourth level of text difficulty. The most amount of variability occurred at the third level.

Clearly, this study is exploratory. Many educational tests are similar to the TEPA, in that sets of items are organized around texts. The application of the bookmark approach seems promising for these types of tests. But when the performance level description is tied to text difficulty as well as item difficulty, item disordinality can cause problems for the judges. It is our hope that the findings from this study will promote further research in this area.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), *Educational Measurement* (pp. 508-600). Washington, DC: American Council on Education.

Lewis, D.M., D.R. Green, H.C. Mitzel, K. Baum, & R.J. Patz (1999). The Bookmark standard setting procedure: Methodology and recent implementations. Manuscript submitted for publication.

Lewis, D.M., H.C. Mitzel, & D.R. Green. (June 1996). *Standard setting: A Bookmark approach*. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Mislevy, R.J. & R.D. Bock. (1989). *BILOG 3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software, Inc.

Pellegrino, J.W., L.R. Jones, & K.J. Mitchell, (eds.) (1999). *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress*. Report of the Committee on the Evaluation of National and State Assessments of Educational Progress, National Research Council. Washington, DC: National Academy Press.

Shepard, L.A. (1995). Implications for standard setting of the National Academy of Education valuation of the National Assessment of Educational Progress achievement levels. Pp. 143-160 in *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments, Volume II*. Presentation at the Joint Conference on Standard Setting for Large-Scale Assessments, October 5. Washington, DC: U.S. Government Printing Office.